Answer Path at NTCIR-7 CCLQA Track

Deng Yu, Xu Bingjing, Liu Song, Wang Cong
Beijing University of Posts and Telecommunications

Abstract

This is the first time that our group participate NTCIR and Answer Path is a brand new system. In this system, we have normally three components as Question Analyzer, Passage Retrieval and Answer Extractor. Question Analyzer used the combination methods of rules and Lucene was the choice of our search engine platform. And in Answer Extraction, we cut the retrieved passage into sentences and utilized Wikipedia resource to sort and evaluate our answers in Biography Question and Definition Question. Other than that, we experimented on clustering method in Event Question, and Relationship Question was treated as the combination of several definition questions. Asides from the main components above, we developed Sentence Resemble Model and Answer Filtering and so on. And there were a lot of components in our plan that would be developed in the future.

1. Introduction

BUPT Apath's entries for 2008 NTCIR-7 conference was first time established and is composed of three main components as Answer Analyzer, Passage Retrieval and Answer Extractor. Answer Analyzer uses a mixture of rule-based and statistics-based method. Passage Retrieval utilizes Lucene toolkit. And in Answer Extractor we make use of Wikipedia resource and single-pass clustering.

The rest of the paper is organized in the following way. Section 2 presents a brief architecture of Answer Path and algorithms utilized in the system. Section 3 describes the result of our system in this evaluation. Section 4 lists all the errors summarized from the answers after comparing with the standard ones, while section 5 raises all the issues and the future work we ought to do. Finally, we make a conclusion in section 6.

2. System and Algorithm

Answer Path is composed of three components: Answer Analyzer, Passage Retrieval and Answer Extractor. In Answer Analyzer, it combines the methods of rules and statistics. We extract keywords from questions to match the patterns of different questions. In Passage Retrieval part, Lucene is the preferred base in the system front for the demand of Chinese search. As to answer extractor, we developed corresponding strategies for different kinds of questions. In Biography, Definition and Relationship questions, we used the same strategy of utilizing the resources of Wikipedia. We used Wikipedia nuggets to evaluate our answer ranking. And in Event questions, we used clustering method to collect the similar events together using the resources of the first paragraph the result pages.

Word segment and named entity recognition is basic technology in Chinese Text Processing. HIT (Harbin Institute of Technology) IR-Lab NLP package is utilized in this system. More than that, Chinese version of Wikipedia and LDC Chinese-English named entity list are used to optimize the results given by HIT system. Wikipedia is an on-line dictionary which mostly gives paraphrase of entities. LDC list gives not only the entity name but also entity type. For instance, "celeb_china" and "celeb_foreign" which used in question classification stand for celebrity. We make use of the two entity list to combine the segment word to named entity further.

In CCLQA, we submitted two runs using different strategies in Passage Retrieval and Answer Extraction for multilingual task. The comparison between the two runs is shown in the following section.

2.1. Question Analysis

For question classification, templates for the four categories: Biography, Definition, Event, and Relationship are defined. And then four scores are calculated for each question, which would be titled with the corresponding category of the highest score. The scoring method is listed below (Full Score is 3.0).

The keyword: name for Biography question, object(s) for Definition or Relationship question, would be extracted in this step. In Relationship Questions we get the keywords using template such as "NE 和 NE 有什么关系", and NE is the keyword in the retrieval.

Table 1. Scoring Rule in Question Classification

	Mark	Score
Biography	Question contains 谁 / 哪位 / 生平 / 事迹	2
	Question contains celebrity name (if no celebrity but common name)	0.9 (0.6)
Definition	Question contains 什么是 / 是什么 / 何谓 / 何为 / 定义	2
	Question contains object	1
Event	The main branch ends with 事件 / 事故 / 案件 / 案 / 运动 / 事变 / 风波 / 活动 / 革命 / 起义 / 战斗 / 战争 / 战役	2

Manuscript received November 15, 2008. This work was supported in part by Beijing Natural Science Foundation (No.4073037), Ministry of Education Doctor Foundation (No.20060013007), and National Natural Science Foundation of China (No. 60496327, No. 60575034).

	Question contains 列举 / 列出 / 罗列 / 举出 / 历数 / 经过 / 历程 / 大事 / 始末 / 进程 / 过程 / 起源 / 源起 / 规模	1
Relation-ship	Question contains 关系 / 相互 / 意义 / 影响 / 态度 / 立场 / 反响 / 反应 / 反馈 / 反馈 / 观点 / 看法	2
	Question contains 相互 / 与 / 跟 / 以及 / 及 / 和 / 同 / 对 / 对于 / 关于 or contains object +verb +object	1

2.2. Passage Retrieval

We built up our passage search engine using Lucene toolkit.

As Lucene is not satisfying in dealing with Chinese segmentation, we add HIT segmentation in the system front.

2.2.1. Lucene-based document retrieval

The classification and the keywords are extracted from questions after the Question Analyzer. We constructed the query using these keywords according to the classification. Different strategies are applied in different questions.

As to Biography Questions and Definition Questions, we try to locate the people and object in Wikipedia resources. And from the content extracted from Wikipedia, we attain the features to the people or object, which we count as information nuggets. For example, as an entry "dateof birth = 2009-8-20", we manage to get 2009-8-20 as a valuable information nugget. As it's not standardized in all the Wikipedia entries, we have to discard a lot of useful information but only acquire the standard information and extend to a search query according to them. In the two runs we submitted, query formation are distinguishing. The first run gathers all the information nuggets into a pool, collects all the keywords into a query, and commits one search. The second run forms a query corresponding to a nugget, finds the best answer according to the nugget. But the times of the search are depending on the number of the nuggets.

For Event Questions, in order to retrieve the documents comprehensively the queries should be expanded. The nouns except named entities and verbs are expended by semantic dictionaries including HowNet and Thesaurus ("同义词词林"). We get synonymous though Thesaurus and related words by HowNet. The extended words are satisfied in the retrieval document but not necessary. To sum up the query has the form "&& (entity || alternname1 || alternname2 || …) && (verb1 && verb2… && noun1 && noun2 …) || (extend1|| extend2 || …)". The notation "&&" indicates that the contents after it must appear in the retrieval document. The notation "||" indicates that the contents after it appearing or not in the retrieval document are both OK.

Relationship Questions are taking the similar method as Biography Questions. The only difference is that there are usually two keywords in Relationship Questions, which are utilized similarly as Biography Question.

After the first step document retrieval, query feedback based on local context analysis [Buckley C 1995] is utilized to improve IR performance.

2.3. Answer Extraction

In this component, we utilize distinguishing strategies to different questions. We will discuss these in the following chapters.

2.3.1. Biography Questions

In this kind of question, we can get the exact people names as keywords.

In the E-C run 1, we put in the keyword into Wikipedia. If we can get the valuable information nuggets from Wikipedia, we use the nuggets to do answer ranking. First of all, we divided the first 500 retrieved documents into sentences. Then we score sentence by sentence. If a sentence contains the whole nugget (that is, all the keywords and synonymous words of others), then it can get a 1.0 score added. Instead, if a sentence only contains the main nugget (that is, all the keywords or their synonymous words), it can get 0.7 added. If neither of the two above conditions is met, then we calculate the similarity between the candidate sentence and the nugget [see 2.3.5 for detail]. If the similarity score is higher than 0.4, it will be added to the final score to the sentence. After all the sentences' scores are specified, normalization and sorting are applied and achieve the final ranking result. However, if we can't find the entry in Wikipedia in the first place, we use the backup approach ---- syntax approach. In our system, we only process four syntactic situations:

1. (name keywords) + "是"+VOB(the objective of the verb), if a sentence meets this rule, a score of 0.9 is added.

2. (name keywords) + "的"+N(noun), if a sentence meets this rule, a score of 0.5 is added.

3. If the name keywords are the VOB of the sentence, a score of 0.3 is added.

4. If the name keywords are the POB (the objective the Preposition) of the sentence, then 0.3 is added the sentence score.

In the E-C run 2, after we acquire all the information nuggets as the first run, to each nugget, we get the first 200 returned documents and turn into sentences. We form three filters to score each sentence and each filter has different weight. The three filters are whole information, keyword information, similarity filters which is similar to the methods in the first run, and we got the sentence with the highest score to be the answer to this nugget. As for the people not in Wikipedia, we took the same syntactic method to the first run.

Comparing the two runs, the differences are mainly set on the approach, but essence. For the first run, it will return more answers, and as a result of better recall ratio, but can't guarantee a high accuracy, and get a difficulty of ridding unbalance. To the contrast, in the second run,

the number of answers is related to the information nuggets of Wikipedia, which caused lack of guarantee of recall rate, but this method has an advantage over accuracy and balancing.

In the two runs above, the parameters are determined by our empirical observation, after comparing the results generated under different values. Take the number of documents retrieved as an example, we chose 500 to guarantee the recall rate at the same time considering the efficiency of the system. And the thresholds in sentence similarity calculating are decided by the algorithm we use, and 0.4 is what, in our head, should be counted as a qualified answer.

2.3.2. Definition Questions

The definition questions are taken care of just like the biography questions except that we change the people name into object name.

2.3.3. Event Questions

First we mark a retrieval document a score. The key point is the relevance between the question and the document. We cluster the documents by single-pass method [K. Hammouda, 2005]. The time complexity of the single-pass method is O(nk), in witch the k is the number of the category. The cost is much lower than the methods such as K-mean cluster. The result of the cluster depends on the order of the document. However the cluster is based on the order of the retrieval document. Because the important resources are in the previous arrangement the single-pass method is suitable. The method is described bellow.

Begin
initialize θ
$w_1 \leftarrow X$ to determine the initial class of the center
do get new X
$j \leftarrow \arg\min_j \| X - w_{j'} \|$
if $\| X - w_{j'} \| < \theta$
then X to w_j
else new center $w_{k+1} \leftarrow X$
return $w_1, w_2, \ldots$
End

In the system, θ is set to be a number between 0.15 and 0.2. We find that when the number of the cluster center is about 6, the clustered documents have the highest level of aggregation. The relevance between the question and cluster is an evidence of answer extraction. The algorithm is:

$$S_{ci} = \frac{N_i}{D_i} \log D_i$$

D_i is the number of the i^{th} category; Ni is the number of the named entities in the i^{th} category. S_{ci} is the relevance between the i^{th} category and the question.

Every document has a score L_k that is given by the Lucene retrieval system.

$$S_k = S_{ci} + L_k \qquad (k = 1, 2, \ldots, M)$$

S_k is the relevance between a document and question.

The next is getting the answer sentences from the document.

For the event list question

$$S_c = \left(\log(1+n) + R\right) f(s)$$

N is the named entities' number. R is the relevance between the words of the question and the sentence. $f(s)$ is the knowledge function. If the sentence has the contents that conform to the knowledge $f(s)$ =1, otherwise $f(s)$=0. Then the total score is

$$S = S_k + S_c$$

2.3.4. Relationship Questions

Relationship and Event question are relatively more complex question. We use the similar Strategy. The mainly difference is that the two keywords of Relationship Question must be in the answer sentence.

2.3.5. Similarity Based on Weighted Edit-Distance

For most definition and biography questions, Wikipedia has the paraphrase for the key word. It is mentioned above that the similarity between sentences return by document retrieval and Wikipedia nugget is used for sorting. The similarity is calculated according advanced edit-distance. As we know, edit-distance is a mature method in Machine Translation to find the closest sentence, which has the smallest syntax distance to the target one. A brief introduction of edit-distance dynamic programming algorithm is shown below [Che Wanxiang 2004]:

Given a target sentence A comprised of words $\{A_1, A_2, \ldots, A_i, \ldots, A_I\}$, and a candidate sentence B comprised of words $\{B_1, B_2, \ldots, B_i, \ldots, B_J\}$,

$$Dis(i,j) = \min \begin{pmatrix} Dis(i-1, j) + W_d, \\ Dis(i, j-1) + W_i, \\ Dis(i-1, j-1) + Dis(A_i, B_j) \end{pmatrix}$$

where W_d denotes the delete operation weight, W_i denotes the insert operation weight, and $Dis(A_i, B_j)$ is the distance between word A_i and B_j (See [Liu qun, 2002] how similarity between two words is computed). $Dis(I, J)$ is the final edit-distance between sentence A and B.

However, in Question Answering, what is important is not syntax similarity but the semantic one to select candidate according the nugget. It's supposed that entity present more semantic information than normal word. Besides, it hurts little if the candidate contains key information but also something redundant. Therefore, we reduce the weight for words insert to the target sentence, and define two kinds of delete weights. If the

number of Wikipedia entities contained by a sentence is N_e, and the one of normal words is N_w, the delete weight of entities and normal words are

$$W_e = \beta / (\beta * N_e + N_w)$$
$$W_w = 1 / (\beta * N_e + N_w)$$

respectively, where W_e denotes weight of Wikipedia entity, and W_w denotes weight of normal words, β is set 3.0 in the experiment.

For each word W_i in target sentence, we find the word in candidate sentence which has the biggest similarity with it, and make the rest continuous words to a special word the distance of which is defined to zero with any word.

2.3.6. Answer Filtering

At present, IR systems are not rewarded for returning multiple instances of a single nugget, and run into the problem of returning "more of the same" [Diane Kelly, ciQA2006]. The redundancy in documents produces many answers which only contain duplicated valuable information. Thus, the objective of answer filtering module is deleting the duplicated answers and making a further filter on initial answers returned by answer extractor.

The most common answer validation and reranking approaches usually relied on external semantic resources or exploiting search engine results. These algorithms only modeled each answer candidate separately and didn't consider the potential valuable information provided by the whole answer candidate set, summarized by J. Ko, 2007. And J. Ko's model estimated the joint probability of correctness of all answer candidates. Here we made an experiment on an algorithm in similar global perspective.

Firstly, input the initial answers sequentially, and use the similarity based on weighted edit-distance to remove the one has high score with one of the former answers.

Secondly, an information value measure is proposed here. We make a supposition that the answer would contain some entities which represent relationship with the keyword in the question, and there are limited and small quantity of relationship between these entities and the keyword. We name those entities related entities. If two answers contain the same entities, it is possible that the two give duplicate information. Based on this supposition, an answer filtering algorithm based on entities statistic was proposed. Because the key entities would occur more frequently than other words in the candidates, we count the frequency of all entities and measure their information value by frequency. This algorithm was showed in detail as following:

1. Take the top a% of initial candidates return by answer extractor as the candidate database.
2. Mark all entity word in the candidate database.
3. Count the frequent of all entity word and score them. Take the top b% of these entities as key entity database K.
4. Given an entity E_i and its ranking percent f_i (if 100 entities in total, and E_i rank at the 3rd place, f_i equals 3%), the information value V_i of E_i is calculated by:

$$V_i = 1 - f_i$$

5. For each candidate answer, accumulate the information value of all entities it contains. Delete this candidate if and only if its value is smaller than threshold and it doesn't contain an entity in K which hasn't occurred in the former answers.

In our system, a=50, b=15, and threshold differs with the category of question.

3. Results

3.1 Description of Runs

The Google Translation API with some simple optimizing rule is utilized to translate English question to Chinese one. It's observed that this API performs well except on some proper noun such as Charles Zhang. And our E-C system is a simple combination of translation and C-C system.

The difference between E-C Run 1 and E-C Run 2 has been explained in 2.3.1. In a word, the E-C Run 1 differs from the E-C Run 2 mainly on the search query extension, answer extraction and ranking module for the biography and definition questions. Besides, some Boolean rules are added in E-C Run1 for event and relationship questions. For all types of questions, E-C Run 1 uses a strategy to control the answer length, which may make a balance between precision and recall, but 02 System doesn't.

E-C runs is submitted earlier than the C-C ones, which had not implement query feedback, special retrieval in headline, answer filtering, and geography knowledge for relationship questions. Both two versions of C-C used the same strategy with E-C Run 1, that is, putting all the information gained from Wikipedia in a pool to form a query to acquire the most related documents about the People or Entity, and then verifying sentence by sentence.

C-C Run 2 uses the same algorithms as C-C Run 1. The difference is that Run 1 uses the novel answer filtering algorithm we have mentioned in 2.3.6 but Run 2 doesn't.

3.2 Result Review

Table 2. Result of APath in NTCIR7

F-Score	E-C Run1	E-C Run2	C-C Run1	C-C Run2
Definition	0.1694	0.1734	0.18	0.1818
Biography	0.1165	0.1567	0.1662	0.1741
Relationship	0.1188	0.1336	0.2067	0.1934
Event	0.0706	0.085	0.1298	0.1317
Overall	0.114	0.1316	0.1702	0.1687

Table 2 lists the results of all runs. Comparing the two runs of C-C, it's known that the Answer Filtering works well on relationship questions, which brought

1.3% improvement on the overall F-Score. The algorithm we proposed is supposed to perform better while the accuracy of the initial answer set rise. However, the answer filtering strategy causes some unexpected decrease on the three other kinds of questions. At present, the filtering algorithm is parametric, and has not considered the noise in Wikipedia entity statistic. That may be the reason of degeneracy on performance.

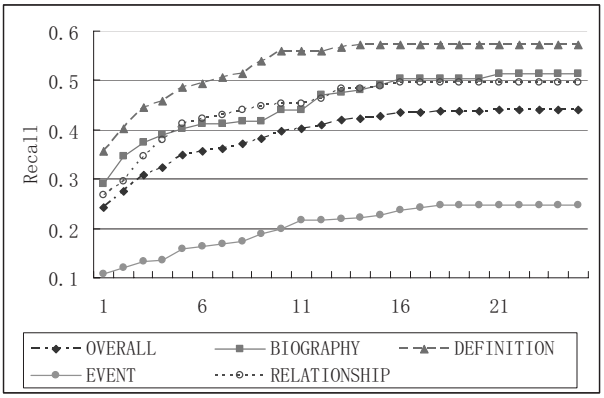

Figure 1: Answer Number - Recall Curve for C-C Run1

Figure 1 reveals that the rising tendency of all recall for four kinds of questions brakes after the system return about 15 sentences. Now it's known that in this corpus overall nugget number is 7.6 and average length is 18.0, which is much smaller than we supposed.

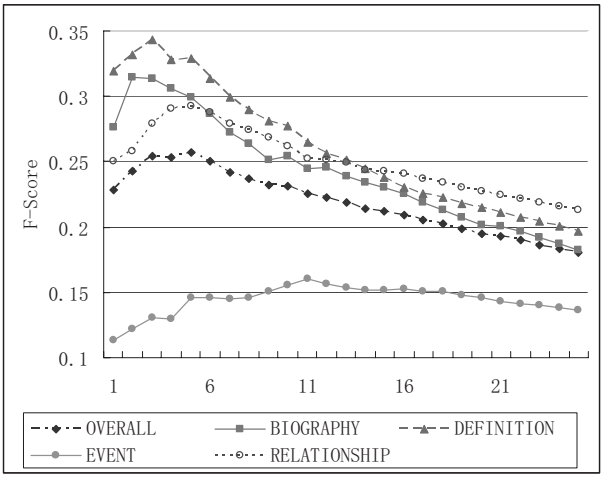

Figure 2: Answer Number – F-Score Curve for C-C Run1

Figure 2 illustrates the curve of F-Score to answer number. F-Score of the system reaches its peak in the 5th answer for overall questions (0.2572), and get the maximum value for Biography questions at (0.3143, 2nd) and Definition questions at (0.3431, 3rd). As for Event and Relationship question, F-Score rises to 0.1605 (in the 15th answer) and 0.2925 (in the 5th answer) respectively.

The major flaw about our system is that the answers we returned are much longer than the length of nugget. The redundant information makes the precision deflating quickly while recall almost stops its increase. This may root in our lack of experience and misunderstanding with the organizers. However, hopefully this will meliorate in the future.

4. Systematic error analysis

After carefully comparing our answers to the standard ones manually, we sum up categories of errors or drawbacks in our system and make suggestions correspondingly which need to be carried out in the future.

In Biography Questions, several kinds errors appears, and the most representative ones are as follows:

(1) Lots of sentences that talks about an event in terms of the people/object, are taken in as answers. For example, in question 317, in English that is "what is aurora?", 10 answers submitted by us including the 4th, 5th, 7th, 13th, 27th are talking about the event that observes an aurora.

(2) Some sentences are describing the people/object, but not in an official way, like 30th answer in Question 317. It's extracted from an essay, but not telling what aurora is.

(3) Other sentences are not mainly about the people/ object which the question asks for, but talk about it incidentally, like the 16th in Question 317. The sentence goes: "作为国家科技攻关项目, 中国极地研究所在那里建立了具有国际先进水平的高空大气综合观测系统,包括电离层, 极光, 地磁, 地面臭氧等 8 台观测仪", which is about the system not the aurora.

(4) Some sentences are talking about the people/ object with the same name but not the same people/ object. Still take Question 317 as an instance, the 22nd answer is about a satellite called "aurora", which is a representative.

(5) Some questions' low performance roots in our methods which highly depend on the quality of Wikipedia content, and Question 381 is greatly influenced.

(6) For some definition question like Question 384, "What is the Big Bang Theory?", the standard questions don't appear Big Bang Theory but universe instead. In this case, we can't retrieve the correct ones because we strictly have the sentence contain the keywords in question.

(7) There are also some issues about the standard answers. Some answers we submitted seem to meet the standard ones are not picked out, some standard ones overlap, and few standard ones in our opinion, can't meet the questions, like Question 349. The question is "什么是货币互换协议?", and the standard one "中泰双方为维护地区金融稳定作出共同努力" can't meet the requirement in our perspective.

In Event and Relationship Questions, the mistakes we have made are in the following:

(1) Miss the essential purpose of the question. In Question 74, "列举中俄之间发生的事情", the events we get are not exactly what the question want , that is , not as import as the standard answers.

(2) Lack of knowledge extension such as words "leader", "all the country", etc which hardly appear in

the correct answers but shows in more definite pattern like “chairman”, “America”.

(3) Lack of understanding the abstract nouns. For example in the question “List the disservice of global warming.”, “Disservice” is an abstract noun. We can get the entity “global warming”. However the abstract nouns may not appear in the document.

There are some general problems too, which appear in nearly in every question. These are:

(1) Short of recall rate. A lot of standard answers don’t appear in our answers, which is the natural defect of our system.

(2) Imperfect of our answer filtering. The main flaws of our answer filtering are parameters which should be determined more reasonably and considering the answers’ length added after the previous filtering, which may descends the precision of the original answers.

(3) All our answers, except for not enough candidate sentences, have the exact number of 30, which was supposed to guarantee the recall rate. But to different questions like Biography Question, the number of answers should be decided by the satisfactory scores of the first few answers, so to guarantee the precision, too.

To sum up, many factors influenced Answer Path’s performance. And the work we still have to work on will be discussed in the following section.

5. Discussion

Answer Path’s performance doesn’t meet our expectation, out of experience, research level and so on, but the experiences we learn from this conference and the communicating are really helpful.

The issues enumerated below are the main problems we encountered during the process of evaluation:

a) Answer number and length

The standard answers themselves are a little confusing. For instance, in topic 42, for “谁是本拉登”, only “恐怖分子” is the correct answer. To be contrast, in topic 379, as for “谁是邓肯”, we got as many as 24 standard answers, including three nick names and two overlapped answers. Actually, we knew organizers’ work was overwhelming, but we think that maybe this can be an issue that we can work on in the future. Other than this, we do think the number of a question should have a standard or at least cover the basic points of the people. In our opinion, if we want to know a person’s biography, five or six nuggets would be the least, but in topic 2, 64, 65, 68, the number of answers is much less than our expecting. It’s suggested to make a guideline for selecting nuggets.

b) Answer Unit

Would a snippet be accepted as legal? As nuggets are fragmentary, a couple of snippets may be the best answer unit which can make the system efficient. However, it is a problem how to generate related fragments, and would it make confusion if the logic relationship between these snippets and question is not so clear. For the convenience and feasibility of evaluation, we finally choose the sentences as our base answer unit and would make explorer on this in the future.

And the following topics are about the work we ought to concentrate on.

In the Event Questions get the meaning of the abstract nouns is most important and difficult. We have to get meaning by reasoning and tendency judging but they are very difficult. We attempt to do some works on this question. We extend the abstract words by semantic dictionary and get several related words. Then the related words are compared with the words of the retrieval documents’ sentence. The similarity of the two set of words is the criteria.

From reference [Andrew Hickl, 2007] we formed an idea of using Wikipedia resources to constitute an object community, which will certainly lead to logic reasoning possible. However, because of time limit, we didn’t get a chance to realize our thoughts. And this will be our future work.

6. Conclusion

This paper describes the main framework of our Q/A system Answer Path. The feature of Apath is that Wikipedia utilization impenetrate in the whole process, and the single-pass method in Event / Relationship questions. Moreover, we proposed similarity based on weighted edit distance for answer candidate scoring, a novel answer filtering algorithm based on statistic. After analyzing and discussion, we find a couple of things we can work on in the future, and hopefully, the next version of Answer Path can be more powerful and intelligent.

References

[1] Diane Kelly, Jimmy Lin. Overview of the TREC 2006: ciQA Task. ACM SIGIR Forum 2006.

[2] Che Wanxiang, Liu Ting, etc. Similar Chinese Sentence Retrieval based on Improved Edit-Distance. Chinese High Technology Letters, 2004

[3] Liu qun, Chinese word similarity computing based on HowNet, Computational Linguistics and Chinese Language Processing, 2002

[4] Buckley C, Salton G, etc. Automatic query expansion using SMART. Proceedings of the 3rd TREC, 1995

[5] K. Hammouda and M. Kamel, “Incremental Document Clustering Using Cluster Similarity Histograms”, The IEEE/WIC International Conference on Web Intelligence (WI 2003), pp: 597-601, 2003.

[6] Andrew Hickl, Kirk Roberts, etc, Question Answering with LCC’s CHAUCER-2 at TREC 2007

[7] Jeongwoo Ko, etc, A Probabilistic Graphial Model for Joint Answer Ranking in Question Answering, ACM SIGIR Forum 2007.